S/N 09/891,080
ART UNIT 2151

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-191135
(43)Date of publication of application : 13.07.1999

(51)Int.Cl. G06K 9/20

(21)Application number : 10-125103
(22)Date of filing : 07.05.1998

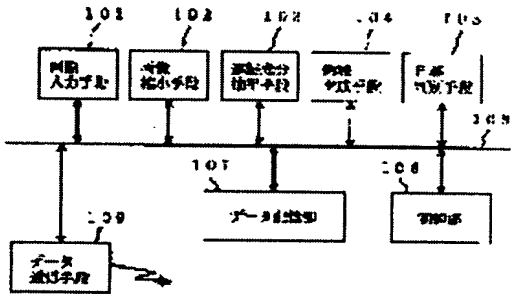
(71)Applicant : RICOH CO LTD
(72)Inventor : MIZUNA TOORU
SAITO TAKASHI

(30)Priority

Priority number : 09245523 Priority date : 10.09.1997 Priority country : JP
09287204 20.10.1997 JP

(54) JAPANESE/ENGLISH DISCRIMINATING METHOD FOR DOCUMENT IMAGE, DOCUMENT RECOGNIZING METHOD AND RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To accurately identify Japanese and English at high speed and to identify both the language even concerning a range to be identified for each character area and each page unit.
SOLUTION: After an input document image 101 is reduced 102, a black pixel connection component is extracted 103, and the character area is generated 104 by merging these components. Concerning the generated character area, based on the length of the connection component, a Japanese/English discriminating means 104 classifies that component and based on the accumulated value of the classified result, whether it is a Japanese area or an English area is discriminated.



LEGAL STATUS

[Date of request for examination] 11.09.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平11－191135

(43)公開日　平成11年(1999)７月13日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｋ　9/20	３２０	Ｇ０６Ｋ　9/20	３２０Ｐ

審査請求　未請求　請求項の数15　ＯＬ　（全 27 頁）

(21)出願番号　特願平10－125103

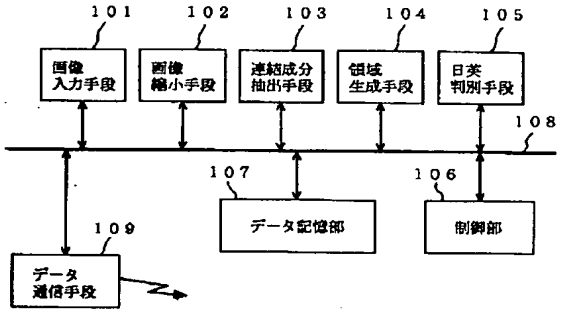
(22)出願日　平成10年(1998)５月７日

(31)優先権主張番号　特願平9－245523
(32)優先日　平９(1997)９月10日
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平9－287204
(32)優先日　平９(1997)10月20日
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　水納　亨
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　齋藤　高志
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　弁理士　鈴木　誠　（外１名）

(54)【発明の名称】　文書画像の日本語英語判定方法、文書認識方法および記録媒体

(57)【要約】
【課題】　精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる。
【解決手段】　入力文書画像１０１を縮小１０２した後、黒画素連結成分を抽出１０３し、それらを統合して文字領域を生成１０４する。生成した文字領域について、日英判別手段１０５は、連結成分の長さを基にその成分を分類し、分類結果の集計値を基に日本語領域であるか英語領域であるかを判別する。



1

【特許請求の範囲】

【請求項１】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴とする文書画像の日本語英語判定方法。

【請求項２】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項３】　前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項２記載の文書画像の日本語英語判定方法。

【請求項４】　各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項５】　ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項６】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項７】　前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項６記載の文書画像の日本語英語判定方法。

【請求項８】　ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像

2

であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項９】　ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１０】　文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項１１】　文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項１２】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１３】　文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１４】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領

域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１５】　文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書画像中の各文字領域に対して日本語領域であるのか英語領域であるのかを判定する文書画像の日本語英語判定方法および記録媒体に関し、また文書画像が日本語文書画像であるか英語文書画像であるかを判定してから認識処理する文書認識方法および記録媒体に関する。

【０００２】

【従来の技術】文書画像に対して文字認識処理を施す場合に、適切な言語を選択する必要がある。すなわち、英文ＯＣＲで日本語を認識しようとしてもアルファベットや数字以外は認識不可能であるし、また逆に日本語ＯＣＲで英文を認識しようとすると、文字切り出しや言語処理のうえで英文ＯＣＲを使用した場合よりも認識率が低くなってしまう。

【０００３】従って、文字認識処理を施す前に、言語識別を行う必要が生じる。従来から文書中の文字種を識別する種々の手法が提案されている。例えば、２値化された文字行の縦方向または横方向の黒白反転回数を計数し、その分布を基に文字種の識別を行う文書認識装置がある（特開平５－１０８８７６号公報を参照）。

【０００４】また、読み取った単語を認識させ、その認識結果と辞書との適合率を基に認識文字の言語種類を判別する文書認識装置もある（特開平６－１５００６１号公報を参照）。

【０００５】

【発明が解決しようとする課題】上記した前者の装置では、文字種を識別する特徴として黒白反転回数を用いているが、この特徴はフォントや文書内容（かな、漢字、数字などの比率）による変動が大きく、このために識別の精度が低くなるという問題がある。

【０００６】これに対して、後者の装置では、一度、文字認識を行っているので、ＯＣＲの性能がよければかなりの確率で字種が判明することになり、精度よく日英判別を行うことが可能となる。しかし、ＯＣＲは処理に多くの時間を要するという問題がある。

【０００７】本発明は上記した事情を考慮してなされたもので、本発明の目的は、精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる文書画像の日本語英語判別方法および記録媒体、さらには、文書画像を判定し、最適な文書認識処理を行う文書認識方法および記録媒体を提供することにある。

【０００８】

【課題を解決するための手段】前記目的を達成するために、請求項１記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴としている。

【０００９】請求項２記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１０】請求項３記載の発明では、前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【００１１】請求項４記載の発明では、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１２】請求項５記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１３】請求項６記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域

であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１４】請求項７記載の発明では、前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【００１５】請求項８記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【００１６】請求項９記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【００１７】請求項１０記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【００１８】請求項１１記載の発明では、文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【００１９】請求項１２記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００２０】請求項１３記載の発明では、文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００２１】請求項１４記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００２２】請求項１５記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００２３】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて具体的に説明する。

〈実施例１〉図１は、本発明の実施例１の構成を示す。図において、１０１は、文書画像を入力する画像入力手段、１０２は、入力文書画像を縮小する画像縮小手段、１０３は、文書画像から連結成分を抽出する連結成分抽出手段、１０４は、抽出した連結成分を分類し、統合することによって文字領域を生成する領域生成手段、１０５は、文字領域単位またはページ単位で日本語と英語を判別する日英判別手段、１０６は、全体を制御する制御部、１０７は、入力された文書画像データや連結成分データ、領域データなど各種データを記憶するデータ記憶部、１０８は、データ通信路、１０９は、ネットワーク、回線などを介してホストなどに接続するデータ通信手段である。

【００２４】図２は、本発明の実施例１の全体の処理フローチャートを示す。以下、図２を参照しながら、本発明の処理動作を説明する。まず、画像入力手段１０１は、文書を読み取ることによって文書画像を得る（ステ

ップ２０１）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段１０９を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【００２５】次に、画像縮小手段１０２は、入力された文書画像を縮小する（ステップ２０２）。この処理は、例えば入力文書画像を１／８程度にＯＲ縮小する処理である。すなわち、８×８画素を１画素に縮小するもので、６４画素中に１つでも黒画素があれば縮小画素は黒画素とする処理である。

【００２６】連結成分抽出手段１０３は、縮小画像から黒画素連結成分を抽出する（ステップ２０３）。領域生成手段１０４は、抽出した連結成分を分類し、統合して文字領域を生成する（ステップ２０４）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された公知の方法を用いればよい。このとき、各文字領域を構成する連結成分の情報はデータ記憶部１０７に格納、保持する。

【００２７】続いて、生成した文字領域について、日英判別手段１０５は日本語か英語かの判定を行う（ステップ２０５）。

【００２８】ステップ２０２において画像をＯＲ縮小することにより、近傍の黒画素どうしが融合する。ここで英文においては単語間にはスペースが存在し、単語内の文字間は非常に狭いという特徴がある。一方、日本語においては、句読点の前後以外では文字間隔は大きくは変わらない。

【００２９】図３は、英文、日本語文の画像例と、その外接矩形を示す。英文画像３０１を縮小し、連結成分を抽出した結果を外接矩形で表現したものが外接矩形３０２である（なお、縮小処理しているので外接矩形３０２は、本来画像３０１より小さくなるべきだが、ここでは同じサイズで表現している）。英文画像では、単語毎に融合して連結成分が構成される。

【００３０】日本語画像３０３と３０５の例について、同様に縮小して連結成分を抽出し、その外接矩形で表現すると、それぞれ外接矩形３０４、３０６のようになる。

【００３１】英文の場合は、単語を構成する文字の数がある程度一定であるので、縦横比が２倍から６、７倍程度となる外接矩形が多くなる特徴がある。一方、日本語の場合は、外接矩形３０４に示すように英文では現れにくい長い矩形が生じたり、逆に外接矩形３０６のように細かい矩形が多く生じる特徴がある。

【００３２】そこで、上記した連結成分矩形を「短」、「中」、「長」の３種類に分類し、これを各文字領域について集計する。図４は、実施例１の日英判定の処理フローチャートを示す。図４の処理は各文字領域毎に行われる。矩形の分類は、行方向が横の場合には例えば、幅／高さが２以下で「短」、幅／高さが２から６で「中」、それ以上で「長」とする（ステップ４０１）。そして、文字領域中におけるこの分類結果を集計し（ステップ４０２）、文字領域毎に日本語か英語かを判定する（ステップ４０３）。ここで、「短」矩形の数をＳＣＮＴ、「中」矩形の数をＮＣＮＴ、「長」矩形の数をＬＣＮＴとすると、日英の判定は図８（ステップ４０３の詳細フローチャート）に示すように行われる。

【００３３】まず、ＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＞Ｔｈ１が成り立つかどうか調べる（ステップ８０１）。Ｔｈ１は予め定めたしきい値であり、例えば０．３程度とする。この条件式が成り立てば、長矩形が十分に多いということであり、当該文字領域は日本語領域であると判定する（ステップ８０４）。

【００３４】次に、ステップ８０１でＮｏと判定されたとき、ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＜Ｔｈ２が成り立つかどうかを調べる（ステップ８０２）。Ｔｈ２も予め定めたしきい値であり、例えば３とする。この条件式が成り立てば、中矩形が少ないということであり、当該文字領域は日本語領域であると判定する（ステップ８０４）。いづれの条件も満たさない場合は、英語領域と判定される（ステップ８０３）。

【００３５】〈実施例２〉上記した実施例１では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないので矩形数の比率で日英判定を行うことが難しくなる可能性がある。実施例２は、矩形の数が十分でない場合を考慮した実施例である。

【００３６】図５は、実施例２の処理フローチャートを示す。日英判別手段１０５は、集計された領域内の矩形の数が十分であるか否か（つまり所定の閾値Ｔｈ以上あるか否か）を調べ（ステップ５０１）、十分でない場合には、前掲した特開平６－１５００６１号公報に記載されているＯＣＲを利用した日英判別を行う（ステップ５０３）。この場合は、文字の数が少ないのでＯＣＲ処理を施しても処理時間の増大は少なくてすむ。そして、矩形の数が十分である場合には実施例１で説明した矩形長による日英の識別を行う（ステップ５０２）。

【００３７】〈実施例３〉次に、ページ単位で日英識別を行う実施例３について説明する。図６、７は、実施例３に係るステップ２０５の詳細フローチャートを示す。図６に示す方法は、「短」、「中」、「長」矩形の数の集計を文字領域毎でなくページ全体について行い（ステップ６０１、６０２）、その結果を使用してページ単位に日英の判定を行う（ステップ６０３）。この日英の判定方法は、図８の処理フローチャートに従って行う。このときのしきい値Ｔｈ１，Ｔｈ２は文字領域単位の処理の場合と異なるしきい値としてもよい。

【００３８】図７に示す方法は、各文字領域毎に日英の判別を行い（ステップ７０２）、その結果を基に当該ペ

ージの日英判定を行う（ステップ７０３）。具体的には、日本語領域と判定された領域の数をＪｎ、英語領域と判定された領域の数をＥｎとして、Ｊｎ＞Ｅｎなら日本語ページ、Ｅｎ＞Ｊｎなら英語ページと判定する。Ｊｎ＝Ｅｎの場合はリジェクトし、あるいは日英の何れかに判定してもよい。

【００３９】〈実施例４〉上記した実施例とは異なる特徴を利用した日英識別方法について説明する。図９は、実施例４の構成を示す。実施例１と異なる点は、行切り出し部９０２と、ブロック抽出部９０３と、ブロック内文字種判別部９０４を設けている点である。他の構成要素は実施例１のものと同様である、図１０は、実施例４の処理フローチャートを示す。

【００４０】まず、行切り出し部９０２は、文書画像の文字領域から行の切り出しを行う（ステップ１００１、１００２）。領域生成処理として、特開平６－２００９２号公報記載の技術を使用した場合には、領域を抽出した段階で行情報が得られているので、これを用いればよく、また電子通信学会論文「周辺密度分布、線密度、外接矩形特徴を利用した文書画像の領域分割」（秋山他、１９８６年8月、Ｖｏｌ．Ｊ６９－Ｄ　Ｎｏ．８）に記載されている射影を用いる方法を用いてもよい。

【００４１】次に、ブロック抽出部９０３は、単語相当のブロックを抽出する（ステップ１００３）。このブロック抽出方法として、本出願人が先に特願平８－３４７８１号で提案した方法を用いればよい。すなわち、ブロック抽出部１１１は、行データ内部の外接矩形を検出し、その外接矩形をブロックデータにまとめる。このブロックデータにまとめる方法は、次の通りである。文字矩形の間隔（まだ一つの矩形が一文字とは確定されていない。従って、漢字の場合、偏とつくりに分離したものがそれぞれ一つの矩形となる場合も多い）のヒストグラムを求める。図１８は、抽出された文字矩形と、矩形間の距離を示す。図１９は、矩形間隔のヒストグラムを示す。

【００４２】このヒストグラムにおいて、最も距離の短いピークは、漢字の偏とつくりの間隔や、プロポーショナル英字の同一単語内の文字間距離に現れる傾向がある。これらを統合しても異なる文字種がブロックに入ることは少ないので、それらを統合することでブロックデータを形成する。この処理を行うことによってプロポーショナルの単語や一文字が分離する（つまり偏とつくりからなる）漢字が一つに統合されることになる。

【００４３】また、最も距離の長いピークは、単語間の距離、句読点と次の文字との距離に現れることが多い。これらは（特に単語間の距離は）文字種が変わる場合の境目に用いられることが多く、同一ブロックになることを避けたい。そこで、最も距離の長いピーク値以上の距離の文字矩形については、同一ブロックにしないように処理する。

【００４４】さらに、対象矩形の両隣の矩形との距離（Ａ，Ｂ）を測定し、その差（Ａ－Ｂ）が所定の閾値以上のとき、長い方の距離の矩形同志は統合せず、短い方の距離の矩形を統合するように処理する。図２０は、矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。図２０では、差が所定の閾値以上大きい位置で矩形の統合を行わないので、３つのブロックが形成される。このような処理を行うことによって、プロポーショナルの英文などで、単語間の距離が絶対的に近くても、文字間距離とは差があるはずであるので、一つの単語だけをまとめて統合できる。また、プロポーショナルフォントであっても日本語の漢字部分は比較的等間隔に配置されるので、日本語文をまとめる場合にも都合がよい。

【００４５】上記したブロック抽出方法を用いることによって、英文の場合、日本語文書と違って単語と単語の間は半角相当のスペースで区切られるために、他の文字種と混合してブロックデータとなることが避けられる。

【００４６】続いて、ブロック内文字種判別部９０４は、ブロック毎の日英判別を行う（ステップ１００４）。これも前掲した出願の方法を用いればよい。つまり、ブロック内文字種判別部９０４は、上記処理によってブロック化されたまとまりが、日本語であるか、英数字であるかという文字種の判定を行う。ブロック内は同一文字種として判断する。この文字種の判定は次のように行う。すなわち、ブロック内の矩形の幅に対して、該矩形の垂直方向の黒ランの数または白黒反転回数が所定の閾値以上のとき日本語文字と識別し、抽出されたブロック内の矩形の垂直方向座標値を基に英字を識別する。図２１（ａ）、（ｂ）は、日本語と英字の場合の垂直方向ランの数の具体例を示す。英数字ではノイズがない理想的な場合、最大で“ｇ”の文字で４つのランができる（図２１（ｂ））。従って、５つ以上のランがカウントされる場合は日本語とする。図２１（ａ）に示す文字「像」の場合、垂直方向のランの数は、文字の下の数字で示すように変化する。

【００４７】日英判別手段９０５は、ブロック毎の判別結果を集計して当該領域の日英判別を行う（ステップ１００５）。ここで、日本語と判定されたブロックの数をＪＣＮＴ、英語と判定されたブロックの数をＥＣＮＴ、不定と判定されたブロックの数をＮＣＮＴとする。図１１は、ステップ１００５の詳細のフローチャートである。ＪＣＮＴ＊Ｔｈ３＞ＥＮＣＴのときは日本語と判定し（ステップ１１０１、１１０５）、そうではなく、ＥＣＮＴ＞ＪＣＮＴのときは英語と判定する（１１０２、１１０４）。それ以外の場合はリジェクトとする（ステップ１１０３）。しきし値Ｔｈ３は、例えば２とする。

【００４８】〈実施例５〉上記した実施例４では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのよ

うな場合は、矩形の数が十分に得られないのでブロックの判別結果数の比率で日英判定を行うことが難しくなる可能性がある。実施例５は、ブロックの数が十分でない場合の実施例である。

【００４９】図１２は、実施例５の処理フローチャートを示す。日英判別手段１０５は、集計された文字領域内のブロックの数が十分であるか否か（つまり所定の閾値Ｔｈ以上あるか否か）を調べ（ステップ１２０１）、十分でない場合には、前掲した特開平６－１５００６１号公報に記載されているＯＣＲを利用した日英判別を行う（ステップ１２０３）。この場合は、文字の数が少ないのでＯＣＲ処理を施しても処理時間の増大は少なくてすむ。そして、ブロックの数が十分である場合には実施例４で説明したブロック毎の判別結果による日英の識別を行う（ステップ１２０２）。

【００５０】〈実施例６〉実施例６は、実施例４の文字領域毎の日英判別を、ページ単位の日英判別に変更したものである。実施例６の処理フローチャートは、図６、７を用いる。

【００５１】図６の処理においては、ＪＣＮＴ、ＥＣＮＴ、ＮＣＮＴの集計を文字領域毎でなくページ全体について行い、その結果を使用して、前述した図１１の処理方法によって日英の判定を行う。このときＴｈ３は文字領域単位の場合とは異なってもよい。

【００５２】図７の処理においては、まず、各文字領域毎に判別し、その結果から当該ページの日英判定を行う。具体的には、日本語領域と判定された領域の数をＪｎ、英語領域と判定された領域の数をＥｎとして、Ｊｎ＞Ｅｎなら日本語ページ、Ｅｎ＞Ｊｎなら英語ページと判定する。Ｊｎ＝Ｅｎの場合はリジェクトとしてもいいし、日英の何れかにしてもよい。

【００５３】〈実施例７〉実施例７では、文字領域毎またはページ単位で日英判別を行う際に、図１３に示すように矩形長を利用する日英判別処理（ステップ１３０１）と、ブロック毎の判別結果を利用する日英判別処理（ステップ１３０２）によって、それぞれ日英の判別を行う。そして、それぞれの判別結果から最終的に日英に判別を行う（ステップ１３０３）。

【００５４】両者共に日本語または英語と判定された場合には、最終結果はそのまま日本語または英語と判定すればよい。何れかがリジェクトと判定された場合には、リジェクトでない方の判定結果を最終結果とする。

【００５５】両者の判定結果が、一方が日本語で、他方が英語で、その結果が一致しない場合には、以下のいづれかの判定をする。

（１）リジェクトとする。

（２）両者の確信度を算出し、値の大きな方の結果を採用する。

矩形長を利用する判別方法の確信度としては、例えば

ＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＞Ｔｈ１で、Ｔｈ１＝０．３の場合にはＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＊２．５の値（ただし上限を１とする）

ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＜Ｔｈ２で、Ｔｈ２＝３の場合には（ＬＣＮＴ＋ＳＣＮＴ）／ＮＣＮＴ＊２．５の値（ただし上限を１とする）

ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＞Ｔｈ２で、Ｔｈ２＝３の場合にはＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＊０．３３の値（ただし上限を１とする）

とする。

【００５６】ブロック毎の判別結果を利用する判別方法の確信度としては、例えば

ＪＣＮＴ＊Ｔｈ３＞ＥＣＮＴで、Ｔｈ３＝２の場合には、ＪＣＴＮ／（ＥＣＮＴ＊３）の値（ただし上限を１とする）

ＥＣＮＴ＞ＪＣＮＴの場合には、ＥＣＮＴ／ＪＣＮＴ＊０．７の値（ただし上限を１とする）

とする。

【００５７】〈実施例８〉図１４は、実施例８の構成を示す。また、図１５は、実施例８の処理フローチャートを示す。この実施例では、入力された文書のページ全体について、日英判別部１４１２は、前述した実施例３、６の方法を用いて、そのページが日本語であるか英語であるかの日英識別処理を行い（ステップ１５０１、１５０２）、その判別結果に基づいて選択部１４０３は英文文書認識部１４０４または日本語文書認識部１４０５を選択し、選択された言語の文書認識処理を行い（ステップ１５０４、１５０５）、その認識結果をディスプレイなどの出力部に出力する（ステップ１５０６）。

【００５８】なお、日本語と英語とではその属性が異なることから、領域分割処理やフォント識別処理なども切り替えた方がよい場合がある。そこで、本実施例の文書認識部は、文字認識処理だけではなく、上記した領域分割処理やフォント識別処理も含まれている。

【００５９】〈実施例９〉図１６は、実施例９の構成を示し、図１７は、実施例９の処理フローチャートを示す。実施例８と異なる点は、日英識別を文字領域毎に行う点である。そのために、領域分割部１６０２は、入力文書を文字領域に分割する（ステップ１７０１、１７０２）。ここで、領域分割部では、日英両方に適応できる領域分割方法を使用する。分割処理された後、日英判別部１６０３は文字領域毎に、例えば前述した実施例１の方法を用いて日英識別処理を行い（ステップ１７０４）、その判別結果に基づいて選択部１６０４は英文文書認識部１６０５または日本語文書認識部１６０６を選択し、選択された言語の文書認識処理を行い（ステップ１７０５、１７０６）、その認識結果をディスプレイなどの出力部１６０７に出力する（ステップ１７０７）。なお、実施例９の文書認識部では、文書認識処理の他にフォント識別処理も行う。

【００６０】〈実施例１０〉前述した各実施例は、黒画

素連結成分や矩形長を特徴量として日本語と英語を判定している。しかし、黒画素連結成分を用いる判定方法は処理時間がかかり、また矩形長を利用する方法はリジェクトの発生が高くなることもある。なお、外接矩形の上辺、下辺の行内での相対位置の頻度分布のピーク位置を基に和文か英文かを識別する方法もあるが（特公平７－２１８１７号公報を参照）、傾きがある文書が入力された場合には、頻度分布が大きく変化し、識別精度が低下してしまうという問題点がある。

【００６１】そこで、本実施例では、行高さに対する、行内の外接矩形の高さのヒストグラムを用いて日本語と英語を識別することにより、文書画像の領域毎に精度よくかつ高速に日本語と英語を識別するものである。そして、上記した日英識別方法でも判別不可能な領域に対しては、別の方法を用いて日英識別を行う。

【００６２】図２２は、実施例１０の構成を示す。また、図２３は、実施例１０の全体の処理フローチャートである。まず、画像入力手段２２０１は、文書を読み取ることによって文書画像を得る（ステップ２３０１）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段２２０７を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【００６３】次に、領域生成手段２２０２は、文字領域を生成する（ステップ２３０２）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された方法を用いればよい。次に、行切り出し手段２２０３は、文字領域から文字認識のための行の切り出しを行なう。つまり、文字の外接矩形を求め、それらを統合して行を生成する（ステップ２３０３）。日英識別手段２２０４は、生成した文字領域について日英識別を行なう（ステップ２３０４）。

【００６４】日英の識別は以下のようにして行う。図２７は、日英識別（ステップ２３０４）の詳細のフローチャートである。図２４は、切り出された行と行内の外接矩形の一例を示す。まず、行高さに対する、行内の外接矩形高さの割合の頻度分布を算出する（ステップ２７０１、２７０２）。行高さをｌｉｎｅｈｅｉｇｈｔ、矩形高さをｈｅｉｇｈｔとする。割合をｈｅｉｇｈｔｒａｔｅ＝ｈｅｉｇｈｔ＊１００／ｌｉｎｅｈｅｉｇｈｔとする。また、図２５のような傾きのある文書の場合は、より精度良く日英識別するために、行高さの代わりにその行の矩形の高さの最大値をｌｉｎｅｈｅｉｇｈｔとして用いてもよい。つまり、傾きのある入力文書については、行内矩形の最大高さに対する、行内各外接矩形高さの割合のヒストグラムを基に日英識別する。

【００６５】上記した割合ｈｅｉｇｈｔｒａｔｅが例えば８０以上の場合の矩形数をｌｃｎｔとし、ｈｅｉｇｈｔｒａｔｅが例えば７０以上８０未満の場合の矩形数をｎｃｎｔとし、ｈｅｉｇｈｔｒａｔｅが例えば４０以上７０未満の場合の矩形数をｓｃｎｔとする。文字領域内のすべての矩形に対し、ｌｃｎｔ，ｎｃｎｔ，ｓｃｎｔを求める。

【００６６】図２６は、日本語文書と英語文書について調べた矩形数の一例を示す。一般に、日本語はｌｃｎｔが大きく、英語はｓｃｎｔが大きいという傾向がある。そこで、所定の閾値ｔｈＪ，ｔｈＥを設定し、ｌｃｎｔ／ｓｃｎｔ＞ｔｈＪのとき日本語と判定し（ステップ２７０３）、ｌｃｎｔ／ｓｃｎｔ＜ｔｈＥのとき英語と判定する（ステップ２７０４）。それ以外のときは不明領域とする（ステップ２７０５）。

【００６７】上記した不明領域に対して、統計的手法を用いて日英識別することができる。図２８は、不明領域に対する詳細な処理フローチャートである。例えば、あらかじめ日本語領域と英語領域の特徴値ｌｃｎｔ、ｎｃｎｔ、ｓｃｎｔを正規化し、その平均値と共分散行列の逆行列を日本語、英語についてそれぞれ求める。そして、平均値と共分散行列の逆行列を用いて、日本語、英語のそれぞれについてマハラノビス距離を求める（ステップ２８０１、２８０２）。

【００６８】日本語のマハラノビス距離をＤｊ、英語のマハラノビス距離をＤｅとするとき、所定の閾値をＭｅ、Ｍｊとすると、Ｄｊ／Ｄｅ＞Ｍｅのとき英語と判定し（ステップ２８０３）、Ｄｊ／Ｄｅ＜Ｍｊのとき日本語と判定する（ステップ２８０４）。何れの条件にも満足しない場合は不明領域と判定する（ステップ２８０５）。なお、上記したマハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【００６９】さらに不明と判定された領域に対して、英文認識の確信度を用いて日英識別を行う。図２９は、ステップ２８０５の詳細な処理フローチャートである。英文認識で確信度を算出する（ステップ２９０１）。次いで、算出された確信度について、例えば６０％以上の確信度をもつ単語の個数をＧｏｏｄ、６０％未満で確信度０でない単語の個数をＢａｄ、確信度が０の単語の個数をＺｅｌｏとする（ステップ２９０２）。

【００７０】日英識別の判定値をＶａｌｕｅとするとき、Ｖａｌｕｅ＝Ｇｏｏｄ／（Ｇｏｏｄ＋Ｂａｄ＋Ｚｅｌｏ）

とし（ステップ２９０３）、Ｖａｌｕｅが所定の閾値ｔｈ　ｅｏｃｒを超えれば（ステップ２９０４）、英語と判定し、それ以下ならば日本語と判定する。

【００７１】なお、Ｚｅｌｏに重み付けしてもよい。Ｚｅｌｏを例えばＢａｄの３個分とすると、Ｖａｌｕｅは、Ｂａｄ＝Ｂａｄ＋Ｚｅｌｏ×３であるから

Ｖａｌｕｅ＝Ｇｏｏｄ／（Ｇｏｏｄ＋Ｂａｄ）

となり、Ｖａｌｕｅが閾値ｔｈ　ｅｏｃｒを超えれば英語、それ以下ならば日本語と判定することもできる。このように、日英識別判定のための文字数が少ない領域でも、英文認識による確信度で日英識別しているので、精

15

度よく領域単位の日英識別が行われる。

【００７２】〈実施例１１〉本実施例は、入力文書画像を縮小した画像から外接矩形を生成し、生成された矩形同士で適当な統合を行い、統合後の矩形長の縦横比のヒストグラムを用いて日英識別をより精度良く行なう実施例である。

【００７３】図３０は、実施例１１の構成を示す。また、図３１は、実施例１１の全体の処理フローチャートである。上記した実施例と同様にして画像入力手段３００１によって入力された文書画像は、画像縮小手段３００２によって縮小される（ステップ３１０１、３１０２）。この処理は、例えば文書画像を１／４程度にＯＲ圧縮（４×４画素を１画素に縮小し、１６画素中に１つでも黒画素があれば縮小画像は黒とする）する。

【００７４】次に、領域生成手段３００３は、文字領域を生成する（ステップ３１０３）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された方法を用いればよい。続いて、矩形統合手段３００４は、日英の特性が良く表れるように、矩形の統合を行なう（ステップ３１０４）。例えば、図３２に示すように、矩形１、２のｙ座標（縦方向）の上下座標が近くかつ、隣同士の矩形１、２のｘ座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。また、例えば、図３３に示すように、左側の矩形１が右側の矩形２をｙ座標で包含する位置関係にありかつ、隣同士の矩形１、２のｘ座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。

【００７５】そして、矩形縦横比（矩形長縦／矩形長横）を用いて、長矩形、中矩形、小矩形、極小矩形の４つの特徴量に分ける（図３４）。一般に、日本語は長矩形の出現する割合が高く、また、英語は中矩形の出現する割合が高い。この特性の違いを利用して、日英識別手段３００５は、識別判定式を作成し、日英識別を行なう（ステップ３１０５）。図３５は、日英識別処理の詳細のフローチャートである。

【００７６】

例えば、領域内での長矩形の領域数ｌｃｎｔ

領域内での中矩形の領域数ｎｃｎｔ

領域内での小矩形の領域数ｓｃｎｔ

領域内での極小矩形の領域数ｓｓｃｎｔ（ノイズの場合が多い）を算出し（ステップ３５０１）、領域内での長矩形の割合ｒａｔｉｏ１＝ｌｃｎｔ／（ｎｃｎｔ＋ｓｃｎｔ）を算出し（ステップ３５０２）、領域内での中矩形の割合ｒａｔｉｏ２＝ｎｃｎｔ／（ｌｃｎｔ＋ｓｃｎｔ）を算出する（ステップ３５０３）。なお、上記割合を算出するとき、ｓｓｃｎｔはノイズとして無視した。

【００７７】そして、ｒａｔｉｏ１をｘ座標、ｒａｔｉｏ２をｙ座標とし、誤識別を極力少なく、日英重なっている部分はリジェクトになるように、日本語領域、英語領域、リジェクト領域に分ける。例えば、ｒａｔｉｏ２／ｒａｔｉｏ１＞ｔｈＥならば英語領域と判定（ステップ３５０４）し、ｒａｔｉｏ２／ｒａｔｉｏ１＜ｔｈＪならば日本語領域と判定し（ステップ３５０５）、それ以外の領域は日英不明とする（ステップ３５０６）。ここで、ｔｈＥ、ｔｈＪは所定の閾値である。

16

【００７８】日英不明と判定された領域に対して、実施例１０と同様に、統計的手法を用いて日英識別する。例えば、あらかじめ日本語領域と英語領域の特徴値ｌｃｎｔ、ｎｃｎｔ、ｓｃｎｔを正規化し、その平均値と共分散行列の逆行列を日本語、英語でそれぞれ求める。平均値と共分散行列の逆行列を用いて日本語、英語のそれぞれのマハラノビス距離を求める。日本語のマハラノビス距離をＤｊ、英語のマハラノビス距離をＤｅとするとき、所定の閾値をＭｅ、Ｍｊとすると、Ｄｊ／Ｄｅ＞Ｍｅのとき英語、Ｄｊ／Ｄｅ＜Ｍｊのとき日本語と判定する。何れの条件も満たさない場合は不明と判定する。なお、マハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【００７９】〈実施例１２〉本発明は上記した実施例に限定されず、ソフトウェアによっても実現することができる。本発明をソフトウェアによって実現する場合には、図３６に示すように、ＣＰＵ、メモリ、表示装置、ハードディスク、キーボード、ＣＤ－ＲＯＭドライブ、スキャナなどからなるコンピュータシステムを用意し、ＣＤ－ＲＯＭなどのコンピュータ読み取り可能な記録媒体には、本発明の日本語英語判定機能、文書認識機能を実現するプログラムなどが記録されている。また、スキャナなどの画像入力手段から入力された文書画像などは一時的にハードディスクなどに格納される。そして、該プログラムが起動されると、一時保存された文書画像データが読み込まれて、日本語英語判定処理、文書認識処理を実行し、その結果をディスプレイなどに出力する。

【００８０】

【発明の効果】以上、説明したように、請求項１、１２記載の発明によれば、複数の判定方法を併用しているので、高精度に日本語と英語とを判別することができる。

【００８１】請求項２、３、６、７、１３、１４記載の発明によれば、文書画像中の文字領域毎に精度よく日本語と英語の判別を行うことができる。

【００８２】請求項４、５、８、９、１３、１４記載の発明によれば、文書画像のページ単位に、精度よく日本語と英語の判別を行うことができる。

【００８３】請求項１０、１１、１５記載の発明によれば、日本語または英語と判定された文書画像に対して、適切な文書認識処理を実行しているので、高精度な認識結果を得ることができる。

【図面の簡単な説明】

【図１】本発明の実施例１の構成を示す。

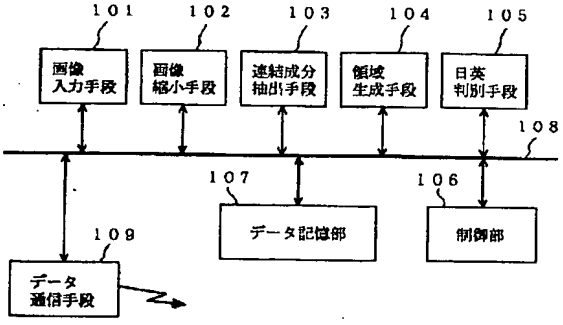
17

【図２】本発明の実施例１の全体の処理フローチャートを示す。
【図３】英文、日本語文の画像例と、その外接矩形を示す。
【図４】実施例１の日英判定の処理フローチャートを示す。
【図５】実施例２の処理フローチャートを示す。
【図６】実施例３に係るステップ２０５の第１の詳細フローチャートを示す。
【図７】実施例３に係るステップ２０５の第２の詳細フローチャートを示す。
【図８】ステップ４０３の詳細フローチャートを示す。
【図９】実施例４の構成を示す。
【図１０】実施例４の処理フローチャートを示す。
【図１１】ステップ１００５の詳細のフローチャートである。
【図１２】実施例５の処理フローチャートを示す。
【図１３】実施例７の処理フローチャートを示す。
【図１４】実施例８の構成を示す。
【図１５】実施例８の処理フローチャートを示す。
【図１６】実施例９の構成を示す。
【図１７】実施例９の処理フローチャートを示す。
【図１８】抽出された文字矩形と、矩形間の距離を示す。
【図１９】矩形間隔のヒストグラムを示す。
【図２０】矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。
【図２１】（ａ）、（ｂ）は、日本語と英字の場合の垂直方向ランの数の具体例を示す。
【図２２】実施例１０の構成を示す。
【図２３】実施例１０の全体の処理フローチャートである。

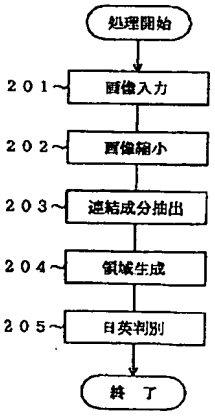
18

＊【図２４】切り出された行と行内の外接矩形の一例を示す。
【図２５】文書が傾いている場合の行と行内の外接矩形の一例を示す。
【図２６】日本語文書と英語文書について調べた矩形数の一例を示す。
【図２７】日英識別（ステップ２３０４）の詳細な処理フローチャートである。
【図２８】不明領域に対する詳細な処理フローチャートである。
【図２９】ステップ２８０５の詳細な処理フローチャートである。
【図３０】実施例１１の構成を示す。
【図３１】実施例１１の全体の処理フローチャートである。
【図３２】矩形を統合する例を示す。
【図３３】矩形を統合する他の例を示す。
【図３４】４種類に分類された矩形を示す。
【図３５】実施例１１の日英識別処理の詳細な処理フローチャートである。
【図３６】実施例１２の構成を示す。
【符号の説明】
１０１　画像入力手段
１０２　画像縮小手段
１０３　連結成分抽出手段
１０４　領域生成手段
１０５　日英判別手段
１０６　制御部
１０７　データ記憶部
１０８　データ通信路
１０９　データ通信手段

【図１】



【図２】



【図３】

This sentence is a sample 301

This sentence is a sample 302

これは、サンプル画像となります。 303

これは、サンプル画像となります。 304

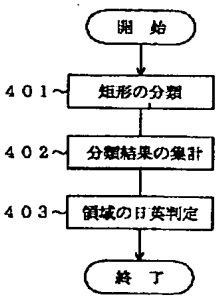
305

これは サ ンプル画像となります。

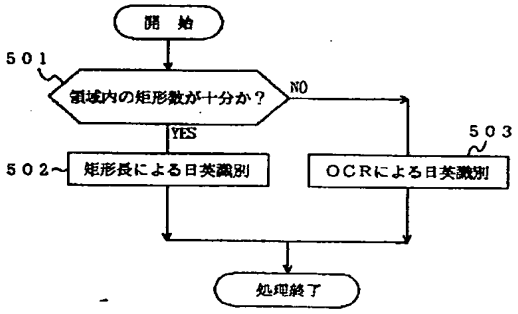
306

これは サンプル画像となります。

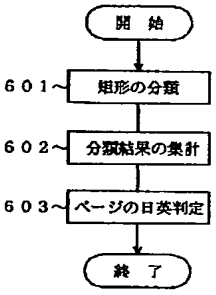
【図４】



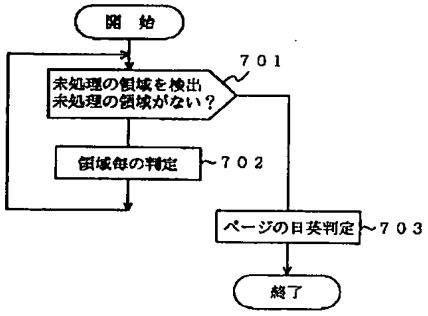
【図５】



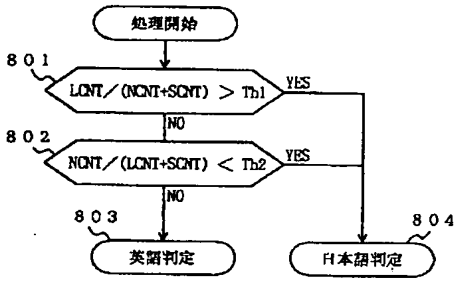
【図６】



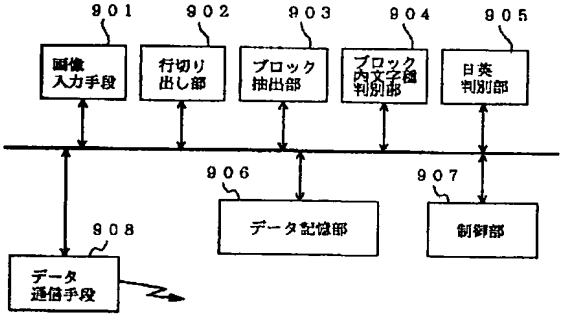
【図７】



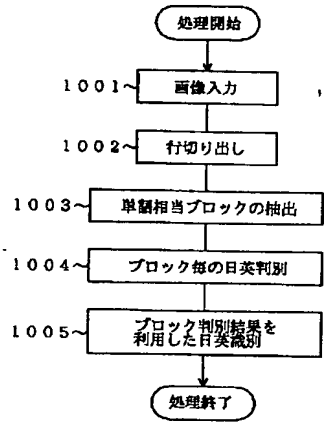
【図８】



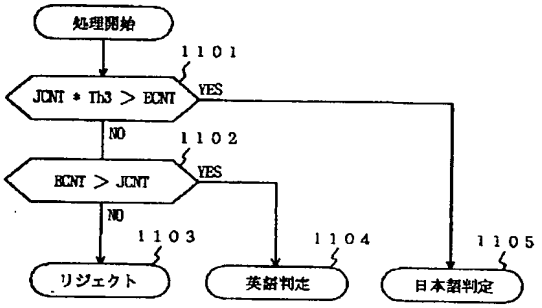
【図９】



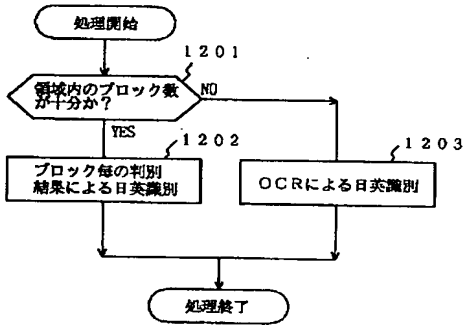
【図１０】



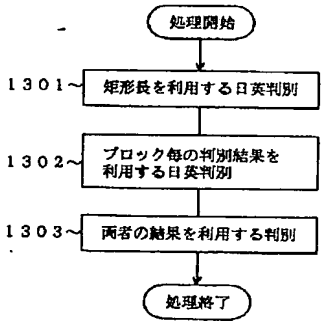
【図１１】



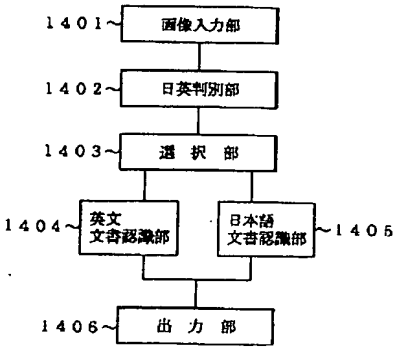
【図１２】



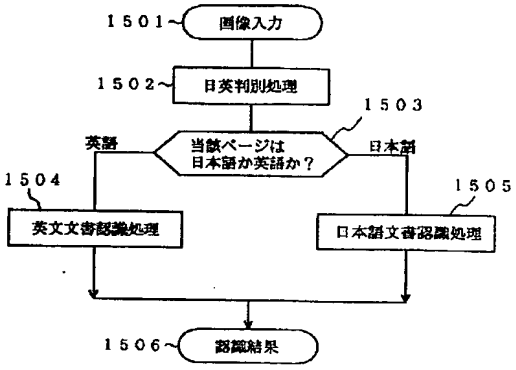
【図１３】



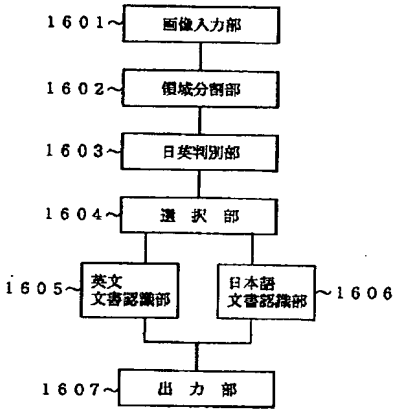
【図１４】



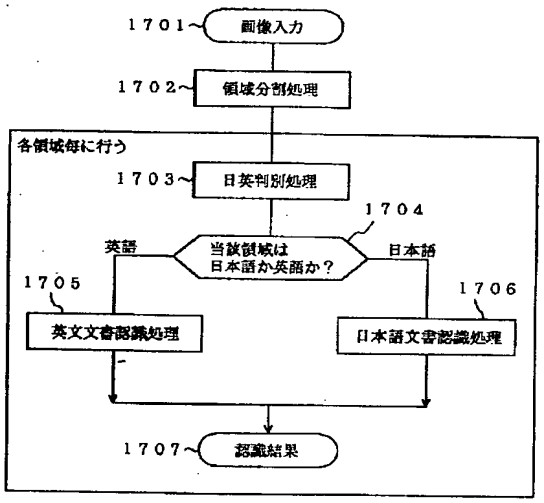
【図15】



【図16】



【図17】

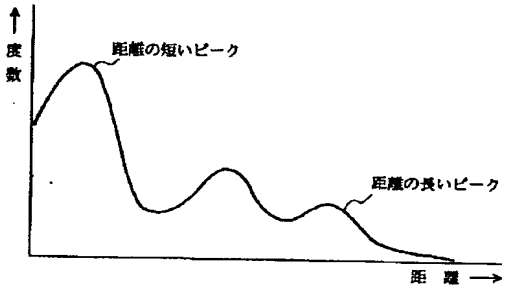


【図18】

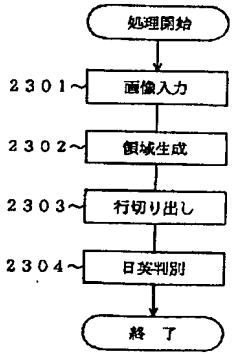
(4) 剛体性を仮定した動画像解析(shape from

矩形間距離

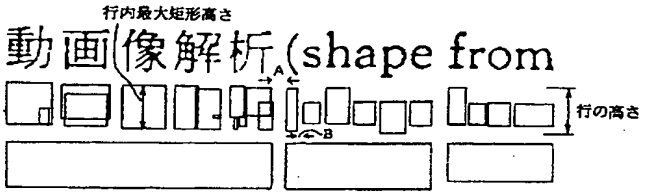
【図19】



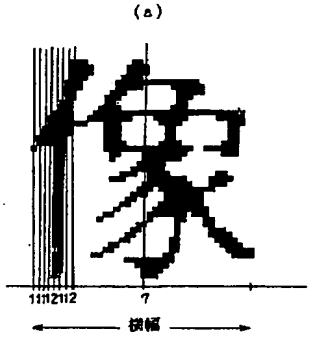
【図23】



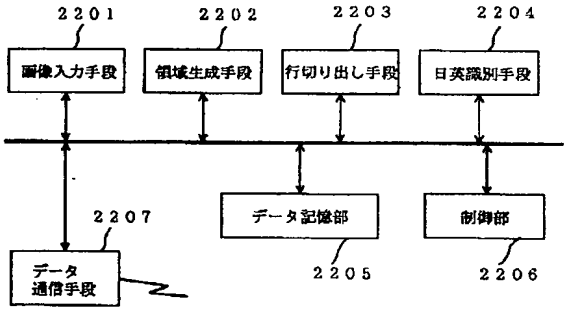
【図２０】



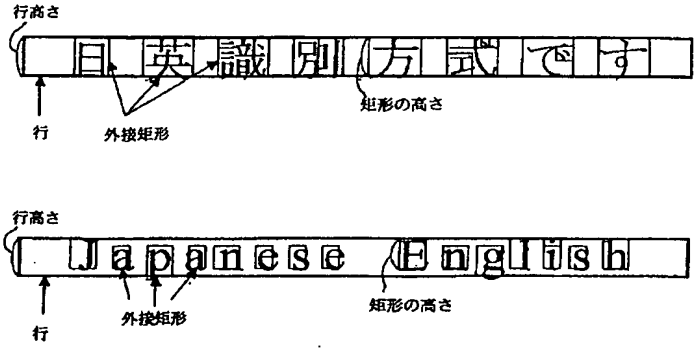
【図２１】



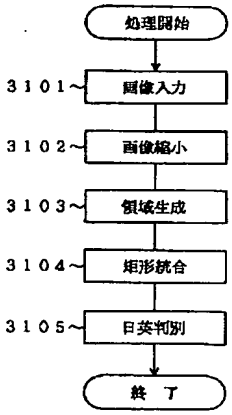
【図２２】



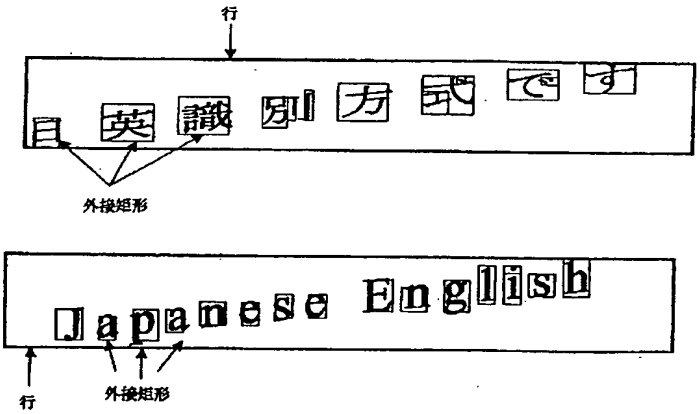
【図２４】



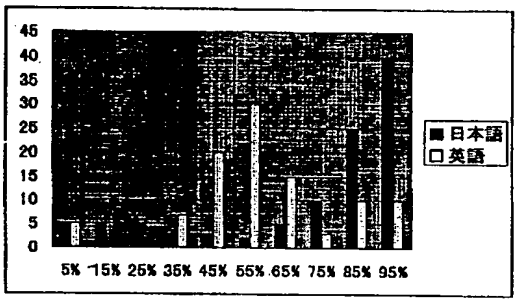
【図３１】



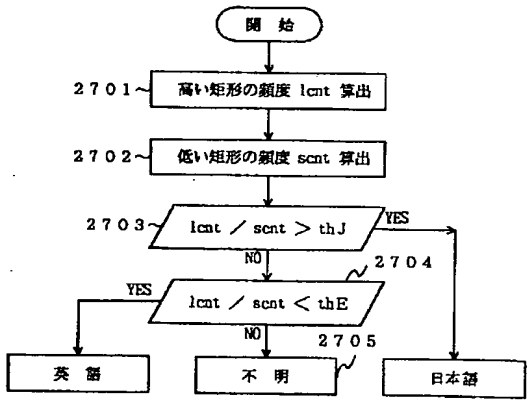
【図25】



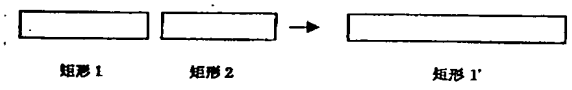
【図26】



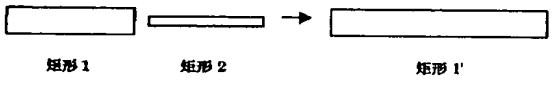
【図27】



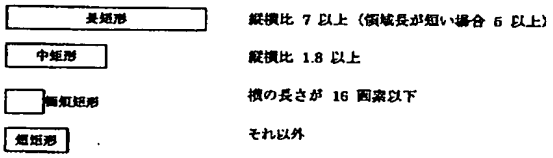
【図32】



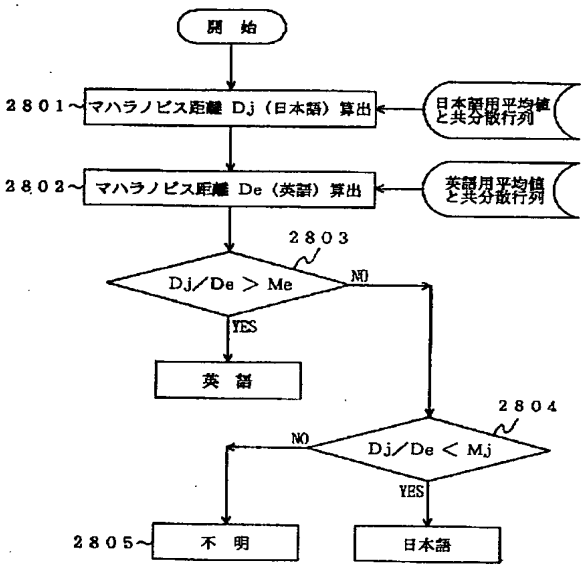
【図33】



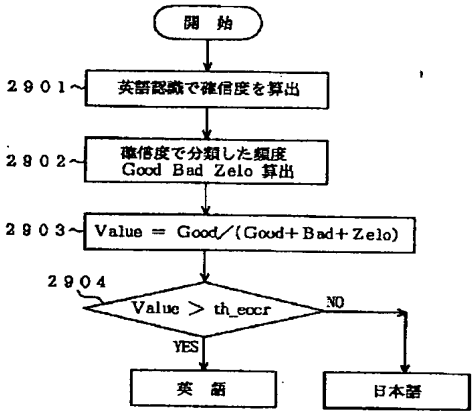
【図34】



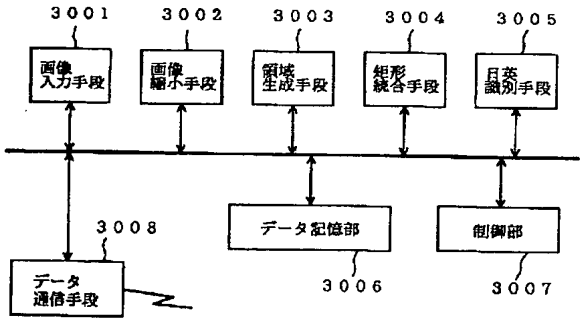
【図２８】



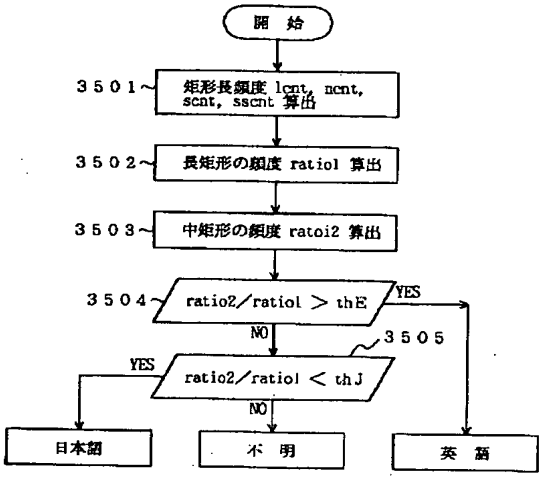
【図２９】



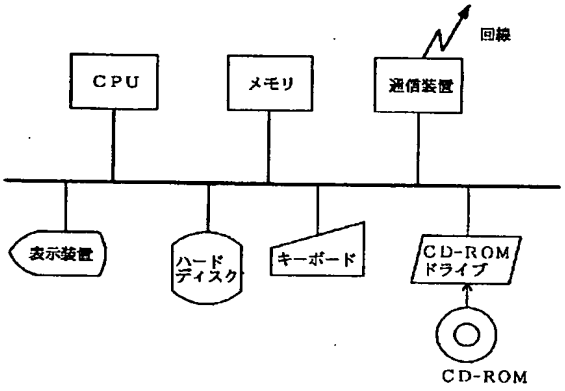
【図３０】



【図３５】



【図３６】



【手続補正書】
【提出日】平成１０年７月１５日
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正内容】
【書類名】　　明細書
【発明の名称】　文書画像の日本語英語判定方法、文書認識方法および記録媒体
【特許請求の範囲】
【請求項１】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴とする文書画像の日本語英語判定方法。
【請求項２】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。
【請求項３】　前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項２記載の文書画像の日本語英語判定方法。
【請求項４】　各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。
【請求項５】　ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。
【請求項６】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。
【請求項７】　前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項６記載の文書画像の日本語英語判定方法。
【請求項８】　ページが複数の文字領域からなり、各ペ

ージの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項９】　ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１０】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１１】　前記行の高さと行内の矩形高さは、行の高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴とする請求項１０記載の文書画像の日本語英語判定方法。

【請求項１２】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１３】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１４】　前記行内の矩形の最大高さと行内の矩形高さは、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴とする請求項１３記載の文書画像の日本語英語判定方法。

【請求項１５】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１６】　前記２度目の判定でも判定できない不明領域については、英文認識の確信度を基に日本語領域であるか英語領域であるかを判定することを特徴とする請求項１２または１５記載の文書画像の日本語英語判定方法。

【請求項１７】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１８】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項１９】　文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項２０】　文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項２１】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを

判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２２】　文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２３】　文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２４】　文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２５】　文書画像中の各文字領域から行を切り出す機能と、行の高さに対する行内の各矩形高さの割合のヒストグラムまたは行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定する機能と、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出する機能と、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２６】　文書画像から縮小画像を生成する機能と、該縮小画像から文字領域を生成する機能と、該文字領域から外接矩形を生成する機能と、所定の位置関係にある矩形同士を統合する機能と、統合後の矩形の縦横比のヒストグラムを基に前記文字領域が日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書画像中の各文字領域に対して日本語領域であるのか英語領域であるのかを判定する文書画像の日本語英語判定方法および記録媒体に関し、また文書画像が日本語文書画像であるか英語文書画像であるかを判定してから認識処理する文書認識方法および記録媒体に関する。

【０００２】

【従来の技術】文書画像に対して文字認識処理を施す場合に、適切な言語を選択する必要がある。すなわち、英文ＯＣＲで日本語を認識しようとしてもアルファベットや数字以外は認識不可能であるし、また逆に日本語ＯＣＲで英文を認識しようとすると、文字切り出しや言語処理のうえで英文ＯＣＲを使用した場合よりも認識率が低くなってしまう。

【０００３】従って、文字認識処理を施す前に、言語識別を行う必要が生じる。従来から文書中の文字種を識別する種々の手法が提案されている。例えば、２値化された文字行の縦方向または横方向の黒白反転回数を計数し、その分布を基に文字種の識別を行う文書認識装置がある（特開平５－１０８８７６号公報を参照）。

【０００４】また、読み取った単語を認識させ、その認識結果と辞書との適合率を基に認識文字の言語種類を判別する文書認識装置もある（特開平６－１５００６１号公報を参照）。

【０００５】

【発明が解決しようとする課題】上記した前者の装置では、文字種を識別する特徴として黒白反転回数を用いているが、この特徴はフォントや文書内容（かな、漢字、数字などの比率）による変動が大きく、このために識別の精度が低くなるという問題がある。

【０００６】これに対して、後者の装置では、一度、文字認識を行っているので、ＯＣＲの性能がよければかなりの確率で字種が判明することになり、精度よく日英判別を行うことが可能となる。しかし、ＯＣＲは処理に多くの時間を要するという問題がある。

【０００７】本発明は上記した事情を考慮してなされたもので、本発明の目的は、精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる文書画像の日本語英語判別方法および記録媒体、さらには、文書画像を判定し、最適な文書認識処理を行う文書認識方法および記録媒体を提供することにある。

【０００８】

【課題を解決するための手段】前記目的を達成するために、請求項１記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴としている。

【０００９】請求項２記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１０】請求項３記載の発明では、前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【００１１】請求項４記載の発明では、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１２】請求項５記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１３】請求項６記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１４】請求項７記載の発明では、前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【００１５】請求項８記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【００１６】請求項９記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【００１７】請求項１０記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００１８】請求項１１記載の発明では、前記行の高さと行内の矩形高さは、行の高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴としている。

【００１９】請求項１２記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２０】請求項１３記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２１】請求項１４記載の発明では、前記行内の矩形の最大高さと行内の矩形高さは、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴としている。

【００２２】請求項１５記載の発明では、文書画像中の

各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２３】請求項１６記載の発明では、前記２度目の判定でも判定できない不明領域については、英文認識の確信度を基に日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２４】請求項１７記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２５】請求項１８記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【００２６】請求項１９記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【００２７】請求項２０記載の発明では、文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【００２８】請求項２１記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００２９】請求項２２記載の発明では、文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００３０】請求項２３記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００３１】請求項２４記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００３２】請求項２５記載の発明では、文書画像中の各文字領域から行を切り出す機能と、行の高さに対する行内の各矩形高さの割合のヒストグラムまたは行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定する機能と、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出する機能と、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００３３】請求項２６記載の発明では、文書画像から

縮小画像を生成する機能と、該縮小画像から文字領域を生成する機能と、該文字領域から外接矩形を生成する機能と、所定の位置関係にある矩形同士を統合する機能と、統合後の矩形の縦横比のヒストグラムを基に前記文字領域が日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【００３４】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて具体的に説明する。

〈実施例１〉図１は、本発明の実施例１の構成を示す。図において、１０１は、文書画像を入力する画像入力手段、１０２は、入力文書画像を縮小する画像縮小手段、１０３は、文書画像から連結成分を抽出する連結成分抽出手段、１０４は、抽出した連結成分を分類し、統合することによって文字領域を生成する領域生成手段、１０５は、文字領域単位またはページ単位で日本語と英語を判別する日英判別手段、１０６は、全体を制御する制御部、１０７は、入力された文書画像データや連結成分データ、領域データなど各種データを記憶するデータ記憶部、１０８は、データ通信路、１０９は、ネットワーク、回線などを介してホストなどに接続するデータ通信手段である。

【００３５】図２は、本発明の実施例１の全体の処理フローチャートを示す。以下、図２を参照しながら、本発明の処理動作を説明する。まず、画像入力手段１０１は、文書を読み取ることによって文書画像を得る（ステップ２０１）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段１０９を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【００３６】次に、画像縮小手段１０２は、入力された文書画像を縮小する（ステップ２０２）。この処理は、例えば入力文書画像を１／８程度にＯＲ縮小する処理である。すなわち、８×８画素を１画素に縮小するもので、６４画素中に１つでも黒画素があれば縮小画素は黒画素とする処理である。

【００３７】連結成分抽出手段１０３は、縮小画像から黒画素連結成分を抽出する（ステップ２０３）。領域生成手段１０４は、抽出した連結成分を分類し、統合して文字領域を生成する（ステップ２０４）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された公知の方法を用いればよい。このとき、各文字領域を構成する連結成分の情報はデータ記憶部１０７に格納、保持する。

【００３８】続いて、生成した文字領域について、日英判別手段１０５は日本語か英語かの判定を行う（ステップ２０５）。

【００３９】ステップ２０２において画像をＯＲ縮小することにより、近傍の黒画素どうしが融合する。ここで英文においては単語間にはスペースが存在し、単語内の文字間は非常に狭いという特徴がある。一方、日本語においては、句読点の前後以外では文字間隔は大きくは変わらない。

【００４０】図３は、英文、日本語文の画像例と、その外接矩形を示す。英文画像３０１を縮小し、連結成分を抽出した結果を外接矩形で表現したものが外接矩形３０２である（なお、縮小処理しているので外接矩形３０２は、本来画像３０１より小さくなるべきだが、ここでは同じサイズで表現している）。英文画像では、単語毎に融合して連結成分が構成される。

【００４１】日本語画像３０３と３０５の例について、同様に縮小して連結成分を抽出し、その外接矩形で表現すると、それぞれ外接矩形３０４、３０６のようになる。

【００４２】英文の場合は、単語を構成する文字の数がある程度一定であるので、縦横比が２倍から６、７倍程度となる外接矩形が多くなる特徴がある。一方、日本語の場合は、外接矩形３０４に示すように英文では現れにくい長い矩形が生じたり、逆に外接矩形３０６のように細かい矩形が多く生じる特徴がある。

【００４３】そこで、上記した連結成分矩形を「短」、「中」、「長」の３種類に分類し、これを各文字領域について集計する。図４は、実施例１の日英判定の処理フローチャートを示す。図４の処理は各文字領域毎に行われる。矩形の分類は、行方向が横の場合には例えば、幅／高さが２以下で「短」、幅／高さが２から６で「中」、それ以上で「長」とする（ステップ４０１）。そして、文字領域中におけるこの分類結果を集計し（ステップ４０２）、文字領域毎に日本語か英語かを判定する（ステップ４０３）。ここで、「短」矩形の数をＳＣＮＴ、「中」矩形の数をＮＣＮＴ、「長」矩形の数をＬＣＮＴとすると、日英の判定は図８（ステップ４０３の詳細フローチャート）に示すように行われる。

【００４４】まず、ＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＞Ｔｈ１が成り立つかどうか調べる（ステップ８０１）。Ｔｈ１は予め定めたしきい値であり、例えば０．３程度とする。この条件式が成り立てば、長矩形が十分に多いということであり、当該文字領域は日本語領域であると判定する（ステップ８０４）。

【００４５】次に、ステップ８０１でＮｏと判定されたとき、ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＜Ｔｈ２が成り立つかどうかを調べる（ステップ８０２）。Ｔｈ２も予め定めたしきい値であり、例えば３とする。この条件式が成り立てば、中矩形が少ないということであり、当該文字領域は日本語領域であると判定する（ステップ８０４）。いづれの条件も満たさない場合は、英語領域と判定される（ステップ８０３）。

【００４６】〈実施例２〉上記した実施例１では、文字

領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないので矩形数の比率で日英判定を行うことが難しくなる可能性がある。実施例２は、矩形の数が十分でない場合を考慮した実施例である。

【００４７】図５は、実施例２の処理フローチャートを示す。日英判別手段１０５は、集計された領域内の矩形の数が十分であるか否か（つまり所定の閾値Ｔｈ以上あるか否か）を調べ（ステップ５０１）、十分でない場合には、前掲した特開平６－１５００６１号公報に記載されているＯＣＲを利用した日英判別を行う（ステップ５０３）。この場合は、文字の数が少ないのでＯＣＲ処理を施しても処理時間の増大は少なくてすむ。そして、矩形の数が十分である場合には実施例１で説明した矩形長による日英の識別を行う（ステップ５０２）。

【００４８】〈実施例３〉次に、ページ単位で日英識別を行う実施例３について説明する。図６、７は、実施例３に係るステップ２０５の詳細フローチャートを示す。図６に示す方法は、「短」、「中」、「長」矩形の数の集計を文字領域毎でなくページ全体について行い（ステップ６０１、６０２）、その結果を使用してページ単位に日英の判定を行う（ステップ６０３）。この日英の判定方法は、図８の処理フローチャートに従って行う。このときのしきい値Ｔｈ１，Ｔｈ２は文字領域単位の処理の場合と異なるしきい値としてもよい。

【００４９】図７に示す方法は、各文字領域毎に日英の判別を行い（ステップ７０２）、その結果を基に当該ページの日英判定を行う（ステップ７０３）。具体的には、日本語領域と判定された領域の数をＪｎ、英語領域と判定された領域の数をＥｎとして、Ｊｎ＞Ｅｎなら日本語ページ、Ｅｎ＞Ｊｎなら英語ページと判定する。Ｊｎ＝Ｅｎの場合はリジェクトし、あるいは日英の何れかに判定してもよい。

【００５０】〈実施例４〉上記した実施例とは異なる特徴を利用した日英識別方法について説明する。図９は、実施例４の構成を示す。実施例１と異なる点は、行切り出し部９０２と、ブロック抽出部９０３と、ブロック内文字種判別部９０４を設けている点である。他の構成要素は実施例１のものと同様である、図１０は、実施例４の処理フローチャートを示す。

【００５１】まず、行切り出し部９０２は、文書画像の文字領域から行の切り出しを行う（ステップ１００１、１００２）。領域生成処理として、特開平６－２００９２号公報記載の技術を使用した場合には、領域を抽出した段階で行情報が得られているので、これを用いればよく、また電子通信学会論文「周辺密度分布、線密度、外接矩形特徴を利用した文書画像の領域分割」（秋山他、１９８６年８月、Ｖｏｌ．Ｊ６９－Ｄ　Ｎｏ．８）に記載されている射影を用いる方法を用いてもよい。

【００５２】次に、ブロック抽出部９０３は、単語相当のブロックを抽出する（ステップ１００３）。このブロック抽出方法として、本出願人が先に特願平８－３４７８１号で提案した方法を用いればよい。すなわち、ブロック抽出部１１１は、行データ内部の外接矩形を検出し、その外接矩形をブロックデータにまとめる。このブロックデータにまとめる方法は、次の通りである。文字矩形の間隔（まだ一つの矩形が一文字とは確定されていない。従って、漢字の場合、偏とつくりに分離したものがそれぞれ一つの矩形となる場合も多い）のヒストグラムを求める。図１８は、抽出された文字矩形と、矩形間の距離を示す。図１９は、矩形間隔のヒストグラムを示す。

【００５３】このヒストグラムにおいて、最も距離の短いピークは、漢字の偏とつくりの間隔や、プロポーショナル英字の同一単語内の文字間距離に現れる傾向がある。これらを統合しても異なる文字種がブロックに入ることは少ないので、それらを統合することでブロックデータを形成する。この処理を行うことによってプロポーショナルの単語や一文字が分離する（つまり偏とつくりからなる）漢字が一つに統合されることになる。

【００５４】また、最も距離の長いピークは、単語間の距離、句読点と次の文字との距離に現れることが多い。これらは（特に単語間の距離は）文字種が変わる場合の境目に用いられることが多く、同一ブロックになることを避けたい。そこで、最も距離の長いピーク値以上の距離の文字矩形については、同一ブロックにしないように処理する。

【００５５】さらに、対象矩形の両隣の矩形との距離（Ａ，Ｂ）を測定し、その差（Ａ－Ｂ）が所定の閾値以上のとき、長い方の距離の矩形同志は統合せず、短い方の距離の矩形を統合するように処理する。図２０は、矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。図２０では、差が所定の閾値以上大きい位置で矩形の統合を行わないので、３つのブロックが形成される。このような処理を行うことによって、プロポーショナルの英文などで、単語間の距離が絶対的に近くても、文字間距離とは差があるはずであるので、一つの単語だけをまとめて統合できる。また、プロポーショナルフォントであっても日本語の漢字部分は比較的等間隔に配置されるので、日本語文をまとめる場合にも都合がよい。

【００５６】上記したブロック抽出方法を用いることによって、英文の場合、日本語文書と違って単語と単語の間は半角相当のスペースで区切られるために、他の文字種と混合してブロックデータとなることが避けられる。

【００５７】続いて、ブロック内文字種判別部９０４は、ブロック毎の日英判別を行う（ステップ１００４）。これも前掲した出願の方法を用いればよい。つまり、ブロック内文字種判別部９０４は、上記処理によっ

てブロック化されたまとまりが、日本語であるか、英数字であるかという文字種の判定を行う。ブロック内は同一文字種として判断する。この文字種の判定は次のように行う。すなわち、ブロック内の矩形の幅に対して、該矩形の垂直方向の黒ランの数または白黒反転回数が所定の閾値以上のとき日本語文字と識別し、抽出されたブロック内の矩形の垂直方向座標値を基に英字を識別する。図２１（ａ）、（ｂ）は、日本語と英字の場合の垂直方向ランの数の具体例を示す。英数字ではノイズがない理想的な場合、最大で“ｇ”の文字で４つのランができる（図２１（ｂ））。従って、５つ以上のランがカウントされる場合は日本語とする。図２１（ａ）に示す文字「像」の場合、垂直方向のランの数は、文字の下の数字で示すように変化する。

【００５８】日英判別手段９０５は、ブロック毎の判別結果を集計して当該領域の日英判別を行う（ステップ１００５）。ここで、日本語と判定されたブロックの数をＪＣＮＴ、英語と判定されたブロックの数をＥＣＮＴ、不定と判定されたブロックの数をＮＣＮＴとする。図１１は、ステップ１００５の詳細のフローチャートである。ＪＣＮＴ＊Ｔｈ３＞ＥＮＣＴのときは日本語と判定し（ステップ１１０１、１１０５）、そうではなく、ＥＣＮＴ＞ＪＣＮＴのときは英語と判定する（１１０２、１１０４）。それ以外の場合はリジェクトとする（ステップ１１０３）。しきし値Ｔｈ３は、例えば２とする。

【００５９】〈実施例５〉上記した実施例４では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないのでブロックの判別結果数の比率で日英判定を行うことが難しくなる可能性がある。実施例５は、ブロックの数が十分でない場合の実施例である。

【００６０】図１２は、実施例５の処理フローチャートを示す。日英判別手段１０５は、集計された文字領域内のブロックの数が十分であるか否か（つまり所定の閾値Ｔｈ以上あるか否か）を調べ（ステップ１２０１）、十分でない場合には、前掲した特開平６－１５００６１号公報に記載されているＯＣＲを利用した日英判別を行う（ステップ１２０３）。この場合は、文字の数が少ないのでＯＣＲ処理を施しても処理時間の増大は少なくてすむ。そして、ブロックの数が十分である場合には実施例４で説明したブロック毎の判別結果による日英の識別を行う（ステップ１２０２）。

【００６１】〈実施例６〉実施例６は、実施例４の文字領域毎の日英判別を、ページ単位の日英判別に変更したものである。実施例６の処理フローチャートは、図６、７を用いる。

【００６２】図６の処理においては、ＪＣＮＴ、ＥＣＮＴ、ＮＣＮＴの集計を文字領域毎でなくページ全体について行い、その結果を使用して、前述した図１１の処理方法によって日英の判定を行う。このときＴｈ３は文字領域単位の場合とは異なってもよい。

【００６３】図７の処理においては、まず、各文字領域毎に判別し、その結果から当該ページの日英判定を行う。具体的には、日本語領域と判定された領域の数をＪｎ、英語領域と判定された領域の数をＥｎとして、Ｊｎ＞Ｅｎなら日本語ページ、Ｅｎ＞Ｊｎなら英語ページと判定する。Ｊｎ＝Ｅｎの場合はリジェクトとしてもいいし、日英の何れかにしてもよい。

【００６４】〈実施例７〉実施例７では、文字領域毎またはページ単位で日英判別を行う際に、図１３に示すように矩形長を利用する日英判別処理（ステップ１３０１）と、ブロック毎の判別結果を利用する日英判別処理（ステップ１３０２）によって、それぞれ日英の判別を行う。そして、それぞれの判別結果から最終的に日英に判別を行う（ステップ１３０３）。

【００６５】両者共に日本語または英語と判定された場合には、最終結果はそのまま日本語または英語と判定すればよい。何れかがリジェクトと判定された場合には、リジェクトでない方の判定結果を最終結果とする。

【００６６】両者の判定結果が、一方が日本語で、他方が英語で、その結果が一致しない場合には、以下のいづれかの判定をする。

（１）リジェクトとする。

（２）両者の確信度を算出し、値の大きな方の結果を採用する。

矩形長を利用する判別方法の確信度としては、例えば

ＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＞Ｔｈ１で、Ｔｈ１＝０．３の場合にはＬＣＮＴ／（ＮＣＮＴ＋ＳＣＮＴ）＊２．５の値（ただし上限を１とする）

ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＜Ｔｈ２で、Ｔｈ２＝３の場合には（ＬＣＮＴ＋ＳＣＮＴ）／ＮＣＮＴ＊２．５の値（ただし上限を１とする）

ＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＞Ｔｈ２で、Ｔｈ２＝３の場合にはＮＣＮＴ／（ＬＣＮＴ＋ＳＣＮＴ）＊０．３３の値（ただし上限を１とする）

とする。

【００６７】ブロック毎の判別結果を利用する判別方法の確信度としては、例えば

ＪＣＮＴ＊Ｔｈ３＞ＥＣＮＴで、Ｔｈ３＝２の場合には、ＪＣＴＮ／（ＥＣＮＴ＊３）の値（ただし上限を１とする）

ＥＣＮＴ＞ＪＣＮＴの場合には、ＥＣＮＴ／ＪＣＮＴ＊０．７の値（ただし上限を１とする）

とする。

【００６８】〈実施例８〉図１４は、実施例８の構成を示す。また、図１５は、実施例８の処理フローチャートを示す。この実施例では、入力された文書のページ全体について、日英判別部１４１２は、前述した実施例３、６の方法を用いて、そのページが日本語であるか英語で

あるかの日英識別処理を行い（ステップ１５０１、１５０２）、その判別結果に基づいて選択部１４０３は英文文書認識部１４０４または日本語文書認識部１４０５を選択し、選択された言語の文書認識処理を行い（ステップ１５０４、１５０５）、その認識結果をディスプレイなどの出力部に出力する（ステップ１５０６）。

【００６９】なお、日本語と英語とではその属性が異なることから、領域分割処理やフォント識別処理なども切り替えた方がよい場合がある。そこで、本実施例の文書認識部は、文字認識処理だけではなく、上記した領域分割処理やフォント識別処理も含まれている。

【００７０】〈実施例９〉図１６は、実施例９の構成を示し、図１７は、実施例９の処理フローチャートを示す。実施例８と異なる点は、日英識別を文字領域毎に行う点である。そのために、領域分割部１６０２は、入力文書を文字領域に分割する（ステップ１７０１、１７０２）。ここで、領域分割部では、日英両方に適応できる領域分割方法を使用する。分割処理された後、日英判別部１６０３は文字領域毎に、例えば前述した実施例１の方法を用いて日英識別処理を行い（ステップ１７０４）、その判別結果に基づいて選択部１６０４は英文文書認識部１６０５または日本語文書認識部１６０６を選択し、選択された言語の文書認識処理を行い（ステップ１７０５、１７０６）、その認識結果をディスプレイなどの出力部１６０７に出力する（ステップ１７０７）。なお、実施例９の文書認識部では、文書認識処理の他にフォント識別処理も行う。

【００７１】〈実施例１０〉前述した各実施例は、黒画素連結成分や矩形長を特徴量として日本語と英語を判定している。しかし、黒画素連結成分を用いる判定方法は処理時間がかかり、また矩形長を利用する方法はリジェクトの発生が高くなることもある。なお、外接矩形の上辺、下辺の行内での相対位置の頻度分布のピーク位置を基に和文か英文かを識別する方法もあるが（特公平７－２１８１７号公報を参照）、傾きがある文書が入力された場合には、頻度分布が大きく変化し、識別精度が低下してしまうという問題点がある。

【００７２】そこで、本実施例では、行高さに対する、行内の外接矩形の高さのヒストグラムを用いて日本語と英語を識別することにより、文書画像の領域毎に精度よくかつ高速に日本語と英語を識別するものである。そして、上記した日英識別方法でも判別不可能な領域に対しては、別の方法を用いて日英識別を行う。

【００７３】図２２は、実施例１０の構成を示す。また、図２３は、実施例１０の全体の処理フローチャートである。まず、画像入力手段２２０１は、文書を読み取ることによって文書画像を得る（ステップ２３０１）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段２２０７を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【００７４】次に、領域生成手段２２０２は、文字領域を生成する（ステップ２３０２）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された方法を用いればよい。次に、行切り出し手段２２０３は、文字領域から文字認識のための行の切り出しを行なう。つまり、文字の外接矩形を求め、それらを統合して行を生成する（ステップ２３０３）。日英識別手段２２０４は、生成した文字領域について日英識別を行なう（ステップ２３０４）。

【００７５】日英の識別は以下のようにして行う。図２７は、日英識別（ステップ２３０４）の詳細のフローチャートである。図２４は、切り出された行と行内の外接矩形の一例を示す。まず、行高さに対する、行内の外接矩形高さの割合の頻度分布を算出する（ステップ２７０１、２７０２）。行高さをｌｉｎｅｈｅｉｇｈｔ、矩形高さをｈｅｉｇｈｔとする。割合をｈｅｉｇｈｔｒａｔｅ＝ｈｅｉｇｈｔ＊１００／ｌｉｎｅｈｅｉｇｈｔとする。また、図２５のような傾きのある文書の場合は、より精度良く日英識別するために、行高さの代わりにその行の矩形の高さの最大値をｌｉｎｅｈｅｉｇｈｔとして用いてもよい。つまり、傾きのある入力文書については、行内矩形の最大高さに対する、行内各外接矩形高さの割合のヒストグラムを基に日英識別する。

【００７６】上記した割合ｈｅｉｇｈｔｒａｔｅが例えば８０以上の場合の矩形数をｌｃｎｔとし、ｈｅｉｇｈｔｒａｔｅが例えば７０以上８０未満の場合の矩形数をｎｃｎｔとし、ｈｅｉｇｈｔｒａｔｅが例えば４０以上７０未満の場合の矩形数をｓｃｎｔとする。文字領域内のすべての矩形に対し、ｌｃｎｔ，ｎｃｎｔ，ｓｃｎｔを求める。

【００７７】図２６は、日本語文書と英語文書について調べた矩形数の一例を示す。一般に、日本語はｌｃｎｔが大きく、英語はｓｃｎｔが大きいという傾向がある。そこで、所定の閾値ｔｈＪ，ｔｈＥを設定し、ｌｃｎｔ／ｓｃｎｔ＞ｔｈＪのとき日本語と判定し（ステップ２７０３）、ｌｃｎｔ／ｓｃｎｔ＜ｔｈＥのとき英語と判定する（ステップ２７０４）。それ以外のときは不明領域とする（ステップ２７０５）。

【００７８】上記した不明領域に対して、統計的手法を用いて日英識別することができる。図２８は、不明領域に対する詳細な処理フローチャートである。例えば、あらかじめ日本語領域と英語領域の特徴値ｌｃｎｔ、ｎｃｎｔ、ｓｃｎｔを正規化し、その平均値と共分散行列の逆行列を日本語、英語についてそれぞれ求める。そして、平均値と共分散行列の逆行列を用いて、日本語、英語のそれぞれについてマハラノビス距離を求める（ステップ２８０１、２８０２）。

【００７９】日本語のマハラノビス距離をＤｊ、英語のマハラノビス距離をＤｅとするとき、所定の閾値をＭｅ、Ｍｊとすると、Ｄｊ／Ｄｅ＞Ｍｅのとき英語と判定

し（ステップ２８０３）、Ｄｊ／Ｄｅ＜Ｍｊのとき日本語と判定する（ステップ２８０４）。何れの条件にも満足しない場合は不明領域と判定する（ステップ２８０５）。なお、上記したマハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【００８０】さらに不明と判定された領域に対して、英文認識の確信度を用いて日英識別を行う。図２９は、ステップ２８０５の詳細な処理フローチャートである。英文認識で確信度を算出する（ステップ２９０１）。次いで、算出された確信度について、例えば６０％以上の確信度をもつ単語の個数をＧｏｏｄ、６０％未満で確信度０でない単語の個数をＢａｄ、確信度が０の単語の個数をＺｅｌｏとする（ステップ２９０２）。

【００８１】日英識別の判定値をＶａｌｕｅとするとき、Ｖａｌｕｅ＝Ｇｏｏｄ／（Ｇｏｏｄ＋Ｂａｄ＋Ｚｅｌｏ）とし（ステップ２９０３）、Ｖａｌｕｅが所定の閾値ｔｈ　ｅｏｃｒを超えれば（ステップ２９０４）、英語と判定し、それ以下ならば日本語と判定する。

【００８２】なお、Ｚｅｌｏに重み付けしてもよい。Ｚｅｌｏを例えばＢａｄの３個分とすると、Ｖａｌｕｅは、Ｂａｄ＝Ｂａｄ＋Ｚｅｌｏ×３であるから

Ｖａｌｕｅ＝Ｇｏｏｄ／（Ｇｏｏｄ＋Ｂａｄ）

となり、Ｖａｌｕｅが閾値ｔｈ　ｅｏｃｒを超えれば英語、それ以下ならば日本語と判定することもできる。このように、日英識別判定のための文字数が少ない領域でも、英文認識による確信度で日英識別しているので、精度よく領域単位の日英識別が行われる。

【００８３】〈実施例１１〉本実施例は、入力文書画像を縮小した画像から外接矩形を生成し、生成された矩形同士で適当な統合を行い、統合後の矩形長の縦横比のヒストグラムを用いて日英識別をより精度良く行なう実施例である。

【００８４】図３０は、実施例１１の構成を示す。また、図３１は、実施例１１の全体の処理フローチャートである。上記した実施例と同様にして画像入力手段３００１によって入力された文書画像は、画像縮小手段３００２によって縮小される（ステップ３１０１、３１０２）。この処理は、例えば文書画像を１／４程度にＯＲ圧縮（４×４画素を１画素に縮小し、１６画素中に１つでも黒画素があれば縮小画像は黒とする）する。

【００８５】次に、領域生成手段３００３は、文字領域を生成する（ステップ３１０３）。この領域生成方法として、例えば特開平６－２００９２号公報に記載された方法を用いればよい。続いて、矩形統合手段３００４は、日英の特性が良く表れるように、矩形の統合を行なう（ステップ３１０４）。例えば、図３２に示すように、矩形１、２のｙ座標（縦方向）の上下座標が近くかつ、隣同士の矩形１、２のｘ座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。また、例えば、図３３に示すように、左側の矩形１が右側の矩形２をｙ座標で包含する位置関係にありかつ、隣同士の矩形１、２のｘ座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。

【００８６】そして、矩形縦横比（矩形長縦／矩形長横）を用いて、長矩形、中矩形、小矩形、極小矩形の４つの特徴量に分ける（図３４）。一般に、日本語は長矩形の出現する割合が高く、また、英語は中矩形の出現する割合が高い。この特性の違いを利用して、日英識別手段３００５は、識別判定式を作成し、日英識別を行なう（ステップ３１０５）。図３５は、日英識別処理の詳細のフローチャートである。

【００８７】

例えば、領域内での長矩形の領域数ｌｃｎｔ

領域内での中矩形の領域数ｎｃｎｔ

領域内での小矩形の領域数ｓｃｎｔ

領域内での極小矩形の領域数ｓｓｃｎｔ（ノイズの場合が多い）を算出し（ステップ３５０１）、領域内での長矩形の割合ｒａｔｉｏ１＝ｌｃｎｔ／（ｎｃｎｔ＋ｓｃｎｔ）を算出し（ステップ３５０２）、領域内での中矩形の割合ｒａｔｉｏ２＝ｎｃｎｔ／（ｌｃｎｔ＋ｓｃｎｔ）を算出する（ステップ３５０３）。なお、上記割合を算出するとき、ｓｓｃｎｔはノイズとして無視した。

【００８８】そして、ｒａｔｉｏ１をｘ座標、ｒａｔｉｏ２をｙ座標とし、誤識別を極力少なく、日英重なっている部分はリジェクトになるように、日本語領域、英語領域、リジェクト領域に分ける。例えば、ｒａｔｉｏ２／ｒａｔｉｏ１＞ｔｈＥならば英語領域と判定（ステップ３５０４）し、ｒａｔｉｏ２／ｒａｔｉｏ１＜ｔｈＪならば日本語領域と判定し（ステップ３５０５）、それ以外の領域は日英不明とする（ステップ３５０６）。ここで、ｔｈＥ、ｔｈＪは所定の閾値である。

【００８９】日英不明と判定された領域に対して、実施例１０と同様に、統計的手法を用いて日英識別する。例えば、あらかじめ日本語領域と英語領域の特徴値ｌｃｎｔ、ｎｃｎｔ、ｓｃｎｔを正規化し、その平均値と共分散行列の逆行列を日本語、英語でそれぞれ求める。平均値と共分散行列の逆行列を用いて日本語、英語のそれぞれのマハラノビス距離を求める。日本語のマハラノビス距離をＤｊ、英語のマハラノビス距離をＤｅとするとき、所定の閾値をＭｅ、Ｍｊとすると、Ｄｊ／Ｄｅ＞Ｍｅのとき英語、Ｄｊ／Ｄｅ＜Ｍｊのとき日本語と判定する。何れの条件も満たさない場合は不明と判定する。なお、マハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【００９０】〈実施例１２〉本発明は上記した実施例に限定されず、ソフトウェアによっても実現することができる。本発明をソフトウェアによって実現する場合に

は、図３６に示すように、ＣＰＵ、メモリ、表示装置、ハードディスク、キーボード、ＣＤ－ＲＯＭドライブ、スキャナなどからなるコンピュータシステムを用意し、ＣＤ－ＲＯＭなどのコンピュータ読み取り可能な記録媒体には、本発明の日本語英語判定機能、文書認識機能を実現するプログラムなどが記録されている。また、スキャナなどの画像入力手段から入力された文書画像などは一時的にハードディスクなどに格納される。そして、該プログラムが起動されると、一時保存された文書画像データが読み込まれて、日本語英語判定処理、文書認識処理を実行し、その結果をディスプレイなどに出力する。

【００９１】

【発明の効果】以上、説明したように、本発明によれば、複数の判定方法を併用しているので、高精度に日本語と英語とを判別することができる。また、文書画像中の文字領域毎に精度よく日本語と英語の判別を行うことができ、文書画像のページ単位に、精度よく日本語と英語の判別を行うことができる。さらに、日本語または英語と判定された文書画像に対して、適切な文書認識処理を実行しているので、高精度な認識結果を得ることができる。

【図面の簡単な説明】

【図１】本発明の実施例１の構成を示す。

【図２】本発明の実施例１の全体の処理フローチャートを示す。

【図３】英文、日本語文の画像例と、その外接矩形を示す。

【図４】実施例１の日英判定の処理フローチャートを示す。

【図５】実施例２の処理フローチャートを示す。

【図６】実施例３に係るステップ２０５の第１の詳細フローチャートを示す。

【図７】実施例３に係るステップ２０５の第２の詳細フローチャートを示す。

【図８】ステップ４０３の詳細フローチャートを示す。

【図９】実施例４の構成を示す。

【図１０】実施例４の処理フローチャートを示す。

【図１１】ステップ１００５の詳細のフローチャートである。

【図１２】実施例５の処理フローチャートを示す。

【図１３】実施例７の処理フローチャートを示す。

【図１４】実施例８の構成を示す。

【図１５】実施例８の処理フローチャートを示す。

【図１６】実施例９の構成を示す。

【図１７】実施例９の処理フローチャートを示す。

【図１８】抽出された文字矩形と、矩形間の距離を示す。

【図１９】矩形間隔のヒストグラムを示す。

【図２０】矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。

【図２１】（ａ）、（ｂ）は、日本語と英字の場合の垂直方向ランの数の具体例を示す。

【図２２】実施例１０の構成を示す。

【図２３】実施例１０の全体の処理フローチャートである。

【図２４】切り出された行と行内の外接矩形の一例を示す。

【図２５】文書が傾いている場合の行と行内の外接矩形の一例を示す。

【図２６】日本語文書と英語文書について調べた矩形数の一例を示す。

【図２７】日英識別（ステップ２３０４）の詳細な処理フローチャートである。

【図２８】不明領域に対する詳細な処理フローチャートである。

【図２９】ステップ２８０５の詳細な処理フローチャートである。

【図３０】実施例１１の構成を示す。

【図３１】実施例１１の全体の処理フローチャートである。

【図３２】矩形を統合する例を示す。

【図３３】矩形を統合する他の例を示す。

【図３４】４種類に分類された矩形を示す。

【図３５】実施例１１の日英識別処理の詳細な処理フローチャートである。

【図３６】実施例１２の構成を示す。

【符号の説明】

１０１　画像入力手段
１０２　画像縮小手段
１０３　連結成分抽出手段
１０４　領域生成手段
１０５　日英判別手段
１０６　制御部
１０７　データ記憶部
１０８　データ通信路
１０９　データ通信手段